Proceedings of LR4SSHOC: Workshop about Language Resources for the SSH Cloud, pages 39–41
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Stretching Disciplinary Boundaries in Language Resource Development and Use: a Linguistic Data Consortium Position Paper

Christopher Cieri
University of Pennsylvania, Linguistic Data Consortium
3600 Market Street, Philadelphia, PA 19104 USA
{ccieri}@ldc.upenn.edu

Abstract

Given the persistent gap between demand and supply, the impetus to reuse language resources is great. Researchers benefit from building upon the work of others including reusing data, tools and methodology. Such reuse should always consider the original intent of the language resource and how that impacts potential reanalysis. When the reuse crosses disciplinary boundaries, the re-user also needs to consider how research standards that differ between social science and humanities on the one hand and human language technologies on the other might lead to differences in unspoken assumptions. Data centers that aim to support multiple research communities have a responsibility to build bridges across disciplinary divides by sharing data in all directions, encouraging re-use and re-sharing and engaging directly in research that improves methodologies.

Keywords: language resources, social sciences and humanities, data centers

1. Introduction

Disciplinary boundaries organize research around shared bodies of knowledge and methods, build consensus, impose order on investigative behavior and create communities of use for purposes of sharing experience (to different degrees in different disciplines). However, given the shortage of Language Resources (LRs), it is frequently necessary to look beyond the traditional disciplinary borders in order to locate data and tools to support modern research. In fact, a deeper look shows that the divisions between academic disciplines have always been porous where the creation of LRs is concerned.

The Linguistic Data Consortium's (LDC) mission since 1992 has been to provide LRs to multiple research communities for purposes of language related education research and technology development. Over that time, LDC has avoided limiting its operations by economic sector, language, geographic region, or academic discipline. LDC supports research communities in three principal ways: 1) publishing corpora from community members to give the data wider use, 2) creating new data sets of value to the community, 3) partnering with members of the research community in new research and providing service via scientific advisory boards, conference program committees and funding panels.

2. Challenges in Data Reuse

Notwithstanding the need and intent to share data across disciplinary boundaries, a potential user of 'found data' must recognize that most corpora have been designed to support specific research agendas. There are exceptions such as the 'national corpora' (e.g. the British National Corpus[1]) that document the state of a language in a specific place and time. However, the focus of data collection effort toward a specific research question may affect its suitability for other uses. A lexicon designed to support machine translation (MT) may contain all the surface forms that appear in a corpus with their glosses into a target language. Another lexicon for the same source language, designed to support speech-to-text (STT) technologies would contain pronunciations rather than glosses. Neither matches the format traditionally used in language teaching where dictionaries are typically organized by a citation form and often contain long form definitions and example sentences in addition to glosses and pronunciations. Could lexicons developed for MT or STT be used in a language teaching situation? Possibly, though that would require either adaptability on the part of the user or adaptation of the LR itself. Student users might find the organization of a dictionary by the actual forms occurring in text more convenient as it removes from them the need to determine the dictionary form. The counter argument that this would cause a ballooning of the size of the dictionary is less important for digital users. Possible augmentations, beyond adding the definitions and example sentences, might include indexing surface forms to citation forms that link to the remainder or the lexical entry. An example of such an approach appears in §5.

This need to enhance a data prior to reuse is not limited to interdisciplinary research (Graff, Bird 2000) describe the long chain of additions, modifications and re-use of two corpora well-known to HLT developers: Switchboard and TDT. The also enumerate the problem that arise when corpus development 'forks' creating multiple versions that are then augmented and used independently.

3. Datasets Created by Social Science and Humanities Researchers

Despite differences in theory, methods and access to resources among the sciences, engineering, social sciences and humanities (SSH), the history of LR development contains multiple example of cross-disciplinary teams and innovative research applying some of the current method of large scale, computational analysis of speech and text among research groups otherwise considered to belong to SSH disciplines.

One of the first publications released by LDC, the HCRC Map Task corpus (LDC93S12), was described as "*a uniquely valuable resource for speech recognition research"* (Anderson, et al. 1991, Thompson et al. 1993) by its creators who described themselves: "*The group which designed and collected the corpus covers a wide range of interests and the corpus reflects this, providing a*

[1] http://www.natcorp.ox.ac.uk/

resource for studies of natural dialogue from many different perspectives." Indeed they worked in research groups named Human Communication, Artificial Intelligence, Cognitive Science, Linguistics and Psychology and were funded by the British Economic and Social Research Council.

Among the ~34 datasets in the LDC datasets that might be called 'lexical', most were designed to support some HLT. However there are several whose intended uses include language teaching or language documentation: Hal Schiffman's English Dictionary of the Tamil Verb (LDC2009L01), Moussa Bamba's dictionaries of Bambara (LDC2016L01), Maninka (LDC2013L01) and Mawu (LDC2005L01), Steven Bird's dictionary of Dschang (LDC2003L01) and Yiwola Awoyale's Global Yoruba Lexical Database (LDC2008L03).

Phoneticians, dialectologists and sociolinguists have also contributed data to LDC in order to reach a broader audience. These include the Digital Archive of Southern Speech - NLP Version (LDC2016S05), the transcribed SLX Corpus of Classic Sociolinguistic Interviews (LDC2003T15) and the Nationwide Speech Project (LDC2007S15) which include words and sentences read under experimental conditions.

4. Reuse of Corpora in SSH

Corpora developed for HLT development have been used successfully in numerous SSH projects. Yaeger-Dror, Hall-Lew and Deckert (2002) select data from numerous publicly available corpora, including 4 from LDC, to correlate negation strategies with dialect, genre and stance. Although the authors were able to build upon the work of many corpus creators, as the paper makes clear, the researcher retains responsibility for understanding the original data, selecting corpora or parts of corpora carefully, augmenting the existing metadata and annotation and anticipating the impact of corpus features on possible analyses. For example, in their analysis of journalistic prose, the authors could draw from many millions of words of news text but decided to select balanced, representative samples of different American regions and match them with other forms of the genre. The news text included bylines but the researchers needed to find the biographies of those writers to determine if they were appropriate exemplars of the dialect regions under study.

5. Research in Social Sciences and Humanities

The use of LRs in language related research, education and technology development has evolved continuously over the past 40 years. Areas of inquiry considered impractical during the US "funding winter" enjoyed a subsequent period significant investment (Liberman 2011, 2015, Church 2017) that continues today and has yielded the successes in multiple HLTs that have in turn enabled their use in SSH research. Others are declared to be solved problems but them subsequently discovered to present unmet challenges (Xu et al. 2019, Cieri et al. 2018, Ryant et al. 2019). The emergence of new tools and methods create opportunities for SSH disciplines to adopt big data approaches. The most efficient of these build upon prior data intensive research including some undertaken outside the discipline. Making connections among research communities to share data and methods is an activity where data centers have a role if not responsibility.

Yuan and Liberman (2008) selected a large sample of US Supreme Court Oral arguments and transcripts provided by the OYEZ[2] project, applied forced-alignment to time-stamp each utterance as to where it occurs in the audio and applied diarization technology to identify the speaker in each case. These technologies increase public access to the deliberations of the court.

Another area where HLT-driven innovation has potential for wide benefit is in language teaching. In Arabic, learning to read presents challenges resulting from the diglossia, dialect diversity, morphological complexity and orthographic features of the language. Digital dictionaries and morphological analysis can offer the learner insights into the language as well as freedom from some of the inefficiencies of traditional study. LDC's Arabic Reading Enhancement Tool (Maamouri 2009) facilitates learner access to standard learner texts through morphological analysis, parsing, digital lexicon and speech synthesis. When learners click on a word in text, that surface form which may be highly inflected or irregular and written without diacritics is indexed to its dictionary form, the relevant dictionary entry is displayed and the word is optionally diacritized and read aloud synthetically.

Other LDC research in SSH disciplines includes work to increase the empirical robustness of assessing film audience engagement. LDC's James Fiumara and Penn Professor of Cinema Studies and English Peter Decherny are co-PIs on "*Measuring Fan Engagement: Finding and Quantifying Text Reuse in Fan Fiction*". The project created the Fan Engagement Meter[3] presenting visualizations of the re-use of text from movie scripts in fan fiction. To date the site covers the Star Wars, The Hobbit/Lord of the Rings and Harry Potter film franchises. The visualization represents the script on the horizontal axis and degree of reuse on the vertical. Hovering over any part of the visualization displays the relevant portion of the text, shaded to show degree of reuse. Researchers can chose to display reuse as a function of exact or fuzzy match and can overlay plots of dialog by character and of sentiment analysis of the script to explore the relations among character, emotion and engagement.

More recent work includes research on the prosodic correlates of sermonic speech in poetry (Mustazza 2019). In this work, the datasets include the text and audio of readings of poetry that have been time aligned and subsequently analyzed for linguistic features that covary with human classification of the style of reading.

6. Conclusion

Linguistic data centers have an obligation to promote the responsible reuse of LRs whether created for HLT or SSH (or other) research and whether used within or across disciplines. Data centers can meet these obligations by engaging with research communities to offer access to existing data, encourage data sharing, document corpus features that affect reuse and take part directly in research that provide proof of concept and improvements to methodology.

[2] https://www.oyez.org/

[3] https://fanengagement.org

7. Bibliographical References

Anderson, Anne H.; Bader, Miles; Bard, Ellen Gurman; Boyle, Elizabeth; Doherty, Gwyneth; Garrod, Simon; Isard, Stephen; Kowtko, Jacqueline; McAllister, Jan; Miller, Jim; Sotillo, Catherine; Thompson, Henry S.; Weinert, Regina (1991) The HCRC Map Task Corpus. *Language and Speech*, 34(4):351-366.

Church, Kenneth (2017) Emerging trends: A tribute to Charles Wayne, *Natural Language Engineering* 24(1): 155-160.

Cieri, Christopher, Mark Liberman, Stephanie Strassel, Denise DiPersio, Jonathan Wright, Andrea Mazzucchi, James Fiumara (2018) From 'Solved Problems' to New Challenges: A Report on LDC Activities. Proc. Language Resources and Evaluation Conference, pp. 3265-3269.

Graff, David, Steven Bird (2000) Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies. Proceedings of the Second International Conference on Language Resources and Evaluation, pp. 427-433, Paris: European Language Resources Association.

Liberman, Mark (2011) Lessons for Reproducible Science from the DARPA Speech and Language Program, presented at the AAAS Workshop: The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer, February 17-21, Washington, DC.

Liberman, Mark (2015) Reproducible Research and the Common Task Method, Simons Foundation Lectures, https://www.simonsfoundation.org/event/reproducible-research-and-the- common-task-method.

Mohammed, Mohamed (2009) LDC Arabic Reading Tools: "Read to Succeed" ACTFL 2009: Arabic SIG Meeting, San Diego, CA, November 21.

Mustazza, Chris (2019) In Search of the Sermonic: Hearing Sonic Genre in Poetry Recordings, presented at Plotting Poetry (and Poetics) 3 - Machiner la poésie (et la poétique) 3, 26-27 Sept. 2019, ATILF, Nancy, France.

Ryant, Neville, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, Mark Liberman (2019) The Second DIHARD Diarization Challenge: Dataset, task, and baselines. In Proceedings Interspeech, September 15–19, 2019, Graz, Austria, pp, 978-982.

Thompson, Henry S., Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands, and Cathy Sotillo. (1993) The HCRC Map Task corpus: natural dialogue for speech recognition. Proceedings of the workshop on Human Language Technology (HLT '93). Association for Computational Linguistics, USA, pages 25–30.

Xu, T., Zhang, H., & Zhang, X. (2019) Joint Training ResCNN-based Voice Activity Detection with Speech Enhancement. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) 1157-1162, IEEE.

Yaeger-Dror, Malcah Lauren Hall-Lew, Sharon Deckert (2002) It's not or isn't it? Using large corpora to determine the influences on contraction strategies. Language Variation and Change 14:79–118.

Yuan, Jiahong, Mark Liberman (2008) Speaker Identification on The Scotus Corpus, Proceedings of Acoustics, pp. 5687-90.